DOCUMENT RESUME

ED 438 313 TM 030 624

AUTHOR Sullivan, Jeremy R.
TITLE A Review of Post-1994 Literature on Whether Statistical Significance Tests Should Be Banned.
PUB DATE 2000-01-29
NOTE 30p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Effect Size; Literature Reviews; *Statistical Significance; *Test Use
IDENTIFIERS *Confidence Intervals (Statistics); *Research Replication

ABSTRACT

This paper summarizes the literature regarding statistical significance testing with an emphasis on: (1) the post-1994 literature in various disciplines; (2) alternatives to statistical significance testing; and (3) literature exploring why researchers have demonstrably failed to be influenced by the 1994 American Psychological Association publication manual's "encouragement" to report effect sizes. Also considered are defenses of statistical significance tests. If the most practical goal for researcher was realized, statistical tests would not be completely banned, but would routinely be supplemented with accurate reports of effect size, confidence intervals, and replicability analyses. (Contains 59 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 438 313

A Review of Post-1994 Literature on Whether Statistical Significance Tests Should be Banned

Jeremy R. Sullivan

Texas A&M University 77843-4225

TM030624

J. Sullivan

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX, January 29, 2000.
The author may be contacted at: "**jrs4605@labs.tamu.edu**".

Abstract

The present paper summarizes the literature regarding statistical significance testing with an emphasis on (a) the post-1994 literature in various disciplines, (b) alternatives to statistical significance testing, and (c) literature exploring why researchers have demonstrably failed to be influenced by the 1994 APA publication manual's "encouragement" (p. 18) to report effect sizes. Also considered are defenses of statistical significance tests.

A Review of Post-1994 Literature on Whether Statistical Significance Tests Should be Banned

Researchers have long placed a premium on the use of statistical significance testing, notwithstanding withering criticisms of many conventional practices as regards statistical inference (e.g., Burdenski, 1999; Carver, 1978; Daniel, 1999; McLean & Ernest, 1999; Meehl, 1978; Morrison & Henkel, 1970; Nix & Barnette, 1999; Thompson, 1993, 1998a, 1998b, 1998c, 1999a, 1999b, 1999d). A series of articles on these issues appeared in recent editions of the American Psychologist (e.g., Cohen, 1990; Kupfersmid, 1988; Rosnow & Rosenthal, 1989). Especially noteworthy are recent articles by Cohen (1994), Kirk (1996), Schmidt (1996), and Thompson (1996).

Indeed, the criticism of statistical testing is growing fierce. For example, Rozeboom (1997) recently argued that:

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students... It is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism... (p. 335)

And Tryon (1998) recently lamented in the American Psychologist,

> The fact that statistical experts and investigators publishing in the best journals cannot consistently

> interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in miniscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless substantial... (p. 796)

Schmidt and Hunter (1997), virulent critics of statistical significance testing, similarly argued that, "Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution" (p. 37, emphasis added).

Criticisms of the statistical significance testing procedure are prevalent, and occur across many scientific disciplines. To be sure, this debate is not an esoteric one left for pure statisticians to resolve; applied psychological, educational, medical, and other social science researchers and even clinicians have taken sides and argued their points cogently (Krantz, 1999; Svyantek & Ekeberg, 1995; Zakzanis, 1998). Indeed, a recent empirical study of four disciplines on a decade-by-decade basis found an exponential increase in criticisms across disciplines of statistical testing practices (Anderson, Burnham, & Thompson, 1999). These criticisms are not only ubiquitous, but also are far from new (see Boring, 1919).

The older commentary eventually led to a very important change in the 1994 APA publication manual: an "encouragement" (p. 18) to always report effect sizes. Yet 11 empirical studies now show that this encouragement has had no effect on the actual reporting practices within either one or two volumes of 23 journals in psychology and education (e.g., Kirk, 1996; Thompson, 1999b; Thompson & Snyder, 1998).

Indeed, the very recently published report of the APA Task Force on Statistical Inference states that effect size should always be reported for all primary results (Wilkinson & the Task Force on Statistical Inference, 1999). Yet the Task Force (1999) itself acknowledged that, "Unfortunately, empirical studies of various journals indicate that the effect size of this [APA publication manual] encouragement has been negligible" (p. 599).

The present paper explores these views in detail with an emphasis on (a) the post-1994 literature in various disciplines, (b) proposed alternatives or supplements to statistical significance testing, and (c) literature exploring why researchers have failed to be influenced by the 1994 APA publication manual's "encouragement" (p. 18) to report effect sizes (cf. Thompson, 1999c). Also considered are defenses of statistical significance testing (cf. Abelson, 1997a; Cortina & Dunlap, 1997; Frick, 1996; Hagen, 1997; Robinson & Levin, 1997). An attempt has been made to keep this discussion conceptually

basic, in the hope that all readers, from novice statistician to seasoned statistical veteran, will find the coverage interesting, enlightening, and accessible.

Defenses of Statistical Significance Testing

A perusal of some of the most popular journals in education and psychology would likely indicate that statistical significance testing has seemingly withstood all of the criticism, as it remains a widely-used analytical tool in these fields (Loftus & Masson, 1994; Shrout, 1997). This section addresses several reasons why statistical testing has weathered the storm, and why many researchers still use statistical significance tests. The reasons covered here include (a) the usefulness of statistical significance testing in making categorical statements and testing ordinal claims; (b) researchers' dissatisfaction with the alternatives to statistical testing; and (c) the argument that statistical testing as originally conceived is a logical and sound method of statistical analysis, and persistent misuse is the fault of the researchers that misuse it rather than an indication of inherent flaws within the method. It is worth noting here that most researchers who advocate the continued use of statistical tests readily acknowledge the limitations of statistical significance testing, yet claim that for some research situations, this is one analysis of choice.

Utility in Testing Ordinal Claims

Ordinal claims are defined as those that do not specify size of effect; they specify only order or direction. Thus, "A is larger than B," and "smoking is positively correlated with lung cancer," are examples of ordinal claims because they provide no information about effect size or strength of association. Frick (1996) noted that "for quantitative claims, null hypothesis testing is not sufficient..., but for ordinal claims it is ideal" (p. 379). According to Abelson (1997b), Frick (1996), and Greenwald, Gonzalez, Harris, and Guthrie (1996), the goal of science is not always determining size of effect; testing ordinal claims (i.e., directional hypotheses) and making categorical statements (i.e., asserting that something important or surprising has occurred) are also important goals of science, goals for which statistical significance testing is well-suited.

Lack of Superior Alternatives

Another argument put forward by fans of statistical tests is that proposed alternative methods, such as effect sizes and confidence intervals (discussed below), are less informative than statistical tests, and are equally vulnerable to widespread misinterpretation (Frick, 1996; Harris, 1997). For example, Harris (1997) stated that statistical significance testing "provides useful information that is not easily gleaned from the

corresponding confidence interval: degree of confidence that we have not made a Type III error and likelihood that our sample result is replicable" (p. 10). Cortina and Dunlap (1997) concluded that statistical tests and proposed alternatives such as confidence intervals each have something equally valuable to contribute to science, and therefore should be used in conjunction with each other.

Misused ≠ Misbegotten

Supporters of statistical significance tests argue vehemently that these methods are not inherently misguided or flawed; rather, years of misuse of this logical, powerful, and potentially useful tool have gradually led to its disrepute (Abelson, 1997b; Cortina & Dunlap, 1997; Frick, 1996; Hagen, 1997). Hagen (1997) expressed this point eloquently:

> The logic of the [statistical test] is elegant, extraordinarily creative, and deeply embedded in our methods of statistical inference. It is unlikely that we will ever be able to divorce ourselves from that logic even if someday we decide that we want to.... The [statistical test] has been misinterpreted and misused for decades. This is our fault, not the fault of the [statistical test].... The logic underlying statistical significance testing has not yet been successfully challenged. (p. 22)

And Abelson (1997b) suggested that we

> Create a list of things that people misuse--for example, oboes, ice skates, band saws, skis, and college educations. Would you be inclined to ban them because people make errors with them? Will we want to ban effect sizes, too, when their misuse escalates? (p. 13)

Finally, Cortina and Dunlap (1997) remind us that careful judgment is required in all areas of science, including statistical analysis, and that the "cure" for misuse and misinterpretation lies not in banning the method, but in improving our education and refining our judgment. Indeed, "mindless application of any procedure causes problems, and discarding a procedure because it has been misapplied ensures the proverbial loss of both baby and bathwater" (Cortina & Dunlap, 1997, p. 171).

Arguments Against Statistical Significance Testing

Several important issues have fueled the arguments against the use of statistical significance tests. Upon reviewing the post-1994 literature, the present author found that the most often-cited and damning issues include those surrounding result replicability, sample size, what statistical significance tests actually tell us, and practical significance. Each of these interrelated issues will be covered separately below, followed

by a discussion of proposed alternatives (or supplements) to the use of statistical significance tests.

The Issue of Replicability

One of the most powerful arguments against the use of statistical significance testing is that these analyses tell neither the researcher nor the research consumer anything about the replicability of a study's results. According to Thompson, the importance of replication in psychological and educational research has enjoyed increased awareness as

> Social scientists have increasingly recognized that the *single* study is inherently governed by subjective passion, that ideology frequently drives even analytic choices, and that the protection against the potentially negative consequences of these passions occurs not from feigned objectivity, but arises in the aggregate across studies from an emphasis on replication. (1994b, p. 157, emphasis in original)

The increased role of replication in educational and psychological research has been accompanied by a growing realization that statistical significance testing has severely limited utility, especially with regard to evaluating the likely replicability of study results (Cohen, 1994; Greenwald et al., 1996; Thompson, 1994b, 1995).

If the purpose of science is formulating generalizable insight based on the cumulation of findings that will generalize under stated conditions, and if the most promising strategies to fulfill this purpose emphasize interpretation based on the estimated likelihood that results will replicate, then statistical significance tests are rendered virtually useless for the underlying purpose of science. Thompson (1994b, 1995) has proposed and elaborated upon the use of several methods that researchers can employ to empirically assess the internal replicability of their research results; these methods include cross-validation, the bootstrap, and the jackknife.

The reason that statistical tests do not evaluate result replicability is that, notwithstanding common misperceptions to the contrary (Cohen, 1994), statistical tests do not test the probability that sample results occur in the population (Carver, 1978). As Thompson (1996) explained,

> Put succinctly, $p_{CALCULATED}$ is *the probability (0 to 1.0) of the sample statistics, given the sample size, and assuming the sample was derived from a population in which the null hypothesis (H_0) is exactly true.* (p. 27, emphasis in original)

In short, statistical tests assume (not test) the population, and test (not assume) the sample results! As Cohen (1994) so clearly explained, this is not what researchers want

to do. But as he also noted, the statistical significance test "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (p. 997).

The Problem of Sample Size

Another problem with statistical significance testing is that it can be circuitous, because to some extent statistical tests evaluate the size of the researcher's sample (Thompson, 1996; Zakzanis, 1998). As researchers increase their sample size, they also increase their chances of obtaining statistically significant results. Thus, as Hays argued almost 20 years ago, "virtually any study can be made to show significant results if one uses enough subjects" (1981, p. 293). And as Thompson more recently explained:

> Because statistical significance tests largely evaluate the size of the researcher's sample, and because researchers already know prior to conducting significance tests whether the sample in hand was large or small, outcomes of these statistical tests do not always yield new insight as a return for the effort invested in conducting the tests. (1995, p. 85)

Thus, one can see how a decision to either reject or not reject the null hypothesis is largely dependent upon the

researcher's sample size. As Thompson (1998b) lamented, "Statistical testing becomes a tautological search for enough participants to achieve statistical significance. If we fail to reject, it is only because we've been too lazy to drag in enough participants" (p. 799). If any given nil null hypothesis can automatically be rejected if we just use a large enough sample, what is the purpose of testing the hypothesis?

Statistical Testing Doesn't Tell Us What We Want to Know

Many researchers feel that an overemphasis on statistical significance testing detracts researchers from the primary purposes and goals of science, such as interpreting research outcomes, theory development, and formulating generalizable insight based on the cumulation of scientific findings (Kirk, 1996; Schmidt, 1996; Thompson, 1995). Thus, statistical significance testing does not usually tell us what we want to know, a point that was touched upon in the section on replicability. Indeed, Kirk (1996) reminded us that "even when a significance test is interpreted correctly, the business of science does not progress as it should" (pp. 753-754). Kirk (1996) went on:

> How far would physics have progressed if their researchers had focused on discovering ordinal relationships? What we want to know is the size of the difference between *A* and *B* and the error

> associated with our estimate; knowing that *A* is greater than *B* is not enough. (p. 754)

Thus, researchers lament that while statistical significance tests may be useful in determining the direction of relationships, we also need to know the strength or magnitude of relationships or differences, and statistical tests are useless in this regard.

Statistical Significance vs. Practical Significance

In addition to the preceding arguments against statistical testing, many researchers are concerned with the ubiquitous practice of equating statistically significant findings with findings that are of practical significance. That is, many researchers present their data such that findings that are found to be statistically significant are also interpreted to be useful, meaningful, or important. Kirk (1996) defined the difference between statistical significance and practical significance nicely: "Statistical significance is concerned with whether a research result is due to chance or sampling variability; practical significance is concerned with whether the result is useful in the real world" (p. 746).

Further, Cohen lamented that

> All psychologists know that *statistically significant* does not mean plain-English significant, but if one reads the literature, one often discovers that a

> finding reported in the Results section studded with asterisks implicitly becomes in the Discussion section highly significant or very highly significant, important, big! (1994, p. 1001, emphasis in original)

As Hubbard (1995) said, "All too often the thicket of ostensibly rigorous significance testing conceals the fact that the research problem is unimportant" (p. 1098).

Because we usually know in advance that the null hypothesis is false, the rejection of a null hypothesis is not very informative or important (Cohen, 1990, 1994; Kirk, 1996; Thompson, 1998b). What are important are measures of the strength of association between the independent and dependent variables and measures of effect size (Cohen, 1994; Kirk, 1996; Snyder & Thompson, 1998; Thompson, 1996, 1999a, 1999b). Support for the reporting of these measures on a routine basis in research journals led to the APA's (1994) "encouragement" (p. 18) to authors to report effect sizes within manuscripts submitted for publication, an issue to which we will turn following a discussion of proposed alternatives or supplements to statistical significance tests.

If Not Statistical Significance Tests, then What?

As Cohen (1994) has noted, "Don't look for a magic alternative to [statistical significance testing], some other

objective mechanical ritual to replace it. It doesn't exist" (p. 1001). So what is the conscientious researcher to do? Critics of statistical significance tests have made several suggestions, with the underlying theme being for researchers to examine and interpret their data carefully and thoroughly, rather than relying solely upon p values in determining which results are important enough to examine further and report in journals. Specific suggestions include the use of effect sizes, confidence intervals, and replicability analyses.

The reporting of effect sizes along with statistical test results in journal articles has been advocated by numerous researchers (e.g., Cohen, 1994; Kirk, 1996; Thompson, 1996, 1999b; Wilkinson & the Task Force on Statistical Inference, 1999; Zakzanis, 1998). Effect sizes are measures of the magnitude of a relationship, difference, or effect, and include variance-accounted-for effect sizes (e.g., R^2, eta^2, $omega^2$) and effect sizes based on standardized differences (e.g., standardized differences in means). Reporting effect sizes for research outcomes enables readers to evaluate the stability of results across studies, and also facilitates the use of meta-analyses in future research. In addition, reporting effect sizes can make research results more understandable, thereby aiding in result interpretation. For detailed information on computing and interpreting effect sizes, the reader is referred to writings by

Kirk (1996), Rosenthal (1996), Rosnow and Rosenthal (1996), Snyder and Lawson (1993), and Snyder and Thompson (1998).

The use of confidence intervals around observed differences or computed effect sizes in research studies has also been recommended by numerous researchers (e.g., Cohen, 1990, 1994; Hunter, 1997; Kirk, 1996; Schmidt, 1996; Wilkinson & the Task Force on Statistical Inference, 1999). Arguments for the use of confidence intervals include: (a) they are easy to compute, requiring no more information than that required for a statistical test; (b) they provide a range of values within which the true effect is likely to lie; (c) they are just as useful as statistical significance tests for deciding whether an observed difference is due to chance or sampling variability; and (d) they facilitate the interpretation of results in terms of practical and useful significance (i.e., whether the results are trivial, useful, or important).

Finally, the limitations of statistical tests point to the importance of either internal or external replicability analyses, which provide valuable information that statistical tests simply cannot (e.g., Cohen, 1994; Levin & Robinson, 1999; Robinson & Levin, 1997; Thompson, 1994b, 1995). While only external analyses invoke true replication, few researchers conduct such analyses due to the immense amount of time and effort that these analyses require. The alternative is internal

replication, which can evaluate the likely replicability of extant study results. Internal replication methods include cross-validation, the jackknife, and the bootstrap, and although these methods are not without their limitations (Levin & Robinson, 1999; Robinson & Levin, 1997), they are certainly preferable to doing nothing at all to evaluate replicability, which is what many people erroneously believing that statistical tests evaluate replicability do (i.e., nothing). For guidance in conducting these analyses, see Thompson (1994b, 1995).

Why Researchers have Ignored the APA's Encouragement

The American Psychological Association has responded to some of the criticisms of statistical significance testing and overreliance on p values by encouraging authors to report effect sizes in articles submitted for publication. This encouragement is found in the 4th edition of the Publication Manual of the American Psychological Association (1994), and reads as follows:

> Neither of the two types of probability values [statistical significance tests] reflects the importance (magnitude) of an effect or the strength of a relationship because both probability values depend on sample size.... You are encouraged to provide effect-size information. (APA, 1994, p. 18)

Despite this encouragement, however, empirical research suggests that little, if any, change has occurred in the reporting

practices in psychological journals (Kirk, 1996; Snyder & Thompson, 1998; Thompson & Snyder, 1997, 1998; Vacha-Haase & Ness, 1999; Vacha-Haase & Nilsson, 1998).

Why has the APA's admonition failed? Schmidt and Hunter (1997) cited stubborn researchers, noting that "changing the beliefs and practices of a lifetime...naturally...provokes resistance" (p. 49). Thompson (1999c) claimed that this policy is overly vague, leaving journal editors uncertain regarding how strictly to enforce the encouragement. Thompson argued further that

> To present an "encouragement" in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, "These myriad requirements count; this encouragement doesn't." (1999c, p. 162)

Thus, authors may minimize the importance of reporting effect sizes, due to the APA's lenient treatment of the policy relative to the rigorous stylistic strictness that characterizes most of the publication manual.

According to Thompson (1999c), "editorial requirements have to change before effect size reporting will become normative" (p. 162). It would appear to be up to journal editors, then, to do more than merely encourage effect size reporting. Indeed, some editors have taken this step and now explicitly require the

reporting of effect sizes along with statistical test results (e.g., Heldref Foundation, 1997; Murphy, 1997; Thompson, 1994a). The APA Task Force on Statistical Inference has also taken a stand, stating that "reporting and interpreting effect sizes...is essential to good research," and that researchers should "always present effect sizes for primary outcomes" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599, emphasis added). Ubiquitous change, however, may not occur until the APA makes effect size reporting a strict requirement in the next edition of the Publication Manual.

Conclusion

The present paper has presented some extreme views on both sides of the statistical significance debate, but perhaps the most practical goal for all is compromise. If this goal were realized, statistical tests would not be completely banned, but would be routinely supplemented with accurate reports of effect size, confidence intervals, and replicability analyses. As Shrout (1997) noted, "Significance testing has become a habit that is difficult to break" (p. 1). Maybe we don't need to completely break this habit, but we do need to practice it more responsibly, in a manner that furthers scientific knowledge. Trying to build a science solely on probability values and ordinal claims is a time- and energy-wasting endeavor, and limits the cumulation of scientific knowledge.

Indeed, notwithstanding the movement away from overemphasis on statistical significance and overreliance on p values, it remains important to understand the flawed logic of those who continue to misuse and misinterpret statistical tests. As Thompson noted:

> We must understand the bad implicit logic of persons who misuse statistical tests if we are to have any hope of persuading them to alter their practices--it will not be sufficient merely to tell researchers not to use statistical tests, or to use them more judiciously. (1996, p. 26)

References

Abelson, R. P. (1997a). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 117-141). Mahwah, NJ: Erlbaum.

Abelson, R. P. (1997b). On the surprising longevity of flogged horses: Why there is a case for the significance test. Psychological Science, 8, 12-15.

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (1999). Null hypothesis testing in ecological studies: Problems, prevalence, and an alternative. Manuscript submitted for publication.

Boring, E. G. (1919). Mathematical vs. scientific importance. Psychological Bulletin, 16, 335-338.

Burdenski, T. (1999, January). A review of the latest literature on whether statistical significance tests should be banned. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED 427 084)

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. Psychological Methods, 2, 161-172.

Daniel, L. G. (1999). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. Research in the Schools, 5(2), 3-14.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? Psychophysiology, 33, 175-183.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52, 15-24.

Harris, R. J. (1997). Significance tests have their place. Psychological Science, 8, 8-11.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart & Winston.

Heldref Foundation. (1997). Guidelines for contributors. Journal of Experimental Education, 65, 95-96.

Hubbard, R. (1995). The earth is highly significantly round ($p < .0001$). American Psychologist, 50, 1098.

Hunter, J. E. (1997). Needed: A ban on the significance test. Psychological Science, 8, 3-7.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. Journal of the American Statistical Association, 94, 1372-1381.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. Educational Psychology Review, 11, 143-155.

Loftus, G. R., & Masson, M. J. (1994). Using confidence intervals in within-subject designs. Psychonomic Bulletin and Review, 1, 476-490.

McLean, J. E., & Ernest, J. M. (1999). The role of statistical significance testing in educational research. Research in the Schools, 5(2), 15-22.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). The significance test controversy. Chicago: Aldine.

Murphy, K. R. (1997). Editorial. Journal of Applied Psychology, 82, 3-5.

Nix, T. W., & Barnette, J. J. (1999). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. Research in the Schools, 5(2), 55-57.

Robinson, D., & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. Educational Researcher, 26(5), 21-26.

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. Journal of Social Service Research, 21(4), 37-59.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. Psychological Methods, 1, 331-340.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 335-392). Mahwah, NJ: Erlbaum.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1, 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 37-64). Mahwah, NJ: Erlbaum.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. Psychological Science, 8, 1-2.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Snyder, P. A., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. School Psychology Quarterly, 13, 335-348.

Svyantek, D. J., & Ekeberg, S. E. (1995). The earth is round (so we can probably get there from here). American Psychologist, 50, 1101.

Thompson, B. (Ed.). (1993). Theme issue: Statistical significance testing in contemporary practice [Special issue]. Journal of Experimental Education, 61(4).

Thompson, B. (1994a). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1994b). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality, 62, 157-176.

Thompson, B. (1995). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. Educational and Psychological Measurement, 55, 84-94.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B. (1998a, April). Five methodology errors in educational research: The pantheon of statistical significance and other faux pas. Invited address presented at the annual meeting of the American Educational Research Association, San Diego. (ERIC Document Reproduction Service No. ED 419 023)

Thompson, B. (1998b). In praise of brilliance: Where that praise really belongs. American Psychologist, 53, 799-800.

Thompson, B. (1998c). Review of What if there were no significance tests? by L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.). Educational and Psychological Measurement, 58, 332-344.

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? Theory and Psychology, 9, 165-181.

Thompson, B. (1999b). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. Exceptional Children, 65, 329-337.

Thompson, B. (1999c). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. Educational Psychology Review, 11, 157-169.

Thompson, B. (1999d). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. Theory and Psychology, 9, 193-199. [Invited address presented at the 1997 annual meeting of the American Psychological Association, Chicago.]

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the *Journal of Experimental Education*. Journal of Experimental Education, 66, 75-83.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *JCD* research articles. Journal of Counseling and Development, 76, 436-441.

Tryon, W. W. (1998). The inscrutable null hypothesis. American Psychologist, 53, 796.

Vacha-Haase, T., & Ness, C. M. (1999). Statistical significance testing as it relates to practice: Use within *Professional Psychology: Research and Practice*. Professional Psychology: Research and Practice, 30, 104-105.

Vacha-Haase, T., & Nilsson, J. E. (1998). Statistical significance reporting: Current trends and uses in *MECD*. Measurement and Evaluation in Counseling and Development, 31, 46-57.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

Zakzanis, K. K. (1998). Brain is related to behavior ($p < .05$). Journal of Clinical and Experimental Neuropsychology, 20, 419-427.